TREAT User Guide, version 1.0

Department of Biomedical Statistics and Informatics, Mayo Clinic
Sept 07 2011

Contents

1. Introduction
2. System Requirements
3. Supported sequencing platforms and file formats
   3.1. Illumina Sequencing
      3.1.1. FASTQ
      3.1.2. BAM
      3.1.3. Called Variants
         3.1.3.1. SNV
         3.1.3.2. INDEL
   3.2 Other sequencing format
4. Installation
5. TREAT workflow pipeline
6. Validation of installed TREAT tool
7. Step by Step Instruction to run TREAT
8. Results Navigation
9. Appendix
10. References
11. Contact Information

Introduction

TREAT, Targeted RE-sequencing Annotation Tool, offers an end-to-end solution for analyzing and interpreting targeted re-sequencing data.

Three Modules of TREAT:

- Sequence alignment
- Variant calling
- Variant annotation and filtering, as well as visualization.

The source code and the executable of TREAT are available for download.
http://ndc.mayo.edu/mayo/research/biostat/stand-alone-packages.cfm

An Amazon Cloud Image of the TREAT is also provided for researchers with no access to local bioinformatics infrastructures (See a separate document on our website: AmazonCloudTutorial.PDF)

System Requirements

To use TREAT, user needs to meet these requirements

1. A CentOS Linux workstation with 4-cores and at least 16 GB of RAM. We do not currently support running it on a Windows platform.
2. ~175 gigabyte (GB) of storage space to download and install TREAT.
3. Additional storage space for user data (~ 1 terabyte (TB) for one flow cell run using Illumina HiSeq2000 as of April 2011)

User Supplied Files:

Illumina sequencing platform: for the Illumina platform, the TREAT accepts 3 different types of input files:

a. FASTQ: to run all 3 modules of TREAT

A FASTQ format is a text-based format which normally uses four lines per sequence.

- Line 1 begins with a '@' character and is followed by a sequence identifier.
- Line 2 is the raw sequence letters (ATGC).
- Line 3 begins with a '+' character and is optional followed by the same sequence identifier again.

- Line 4 encodes the quality values for the sequence in Line 2

```
@SEQ_ID
TCAGTGTTACCCAGGCTAGACTGCAGTTGCACAATCTCTGCTCACTGCAAGCTCTACCTCCCTGGTTCAAGTGATTCTCCTGCCT
+SEQ_ID
\daddcc`\_abbb_[\[V]]]^U^Q\YQRXW\Q\[V]]]dad_c\\_^Y_ccccXaYTabaabc`]c^aBBBBBBBBBBBBBBB
```

b. BAM: to run variant calling and annotation modules only

BAM is compressed and binary format file which includes sequences aligned to reference genome. Because it is binary and compressed it requires less storage space. Indexing of BAM aims to achieve fast retrieval of alignment overlapping a specified region without going through the whole alignment.

c. Called Variants: to run variant annotation module only

Called variants include SNVs and INDELs, following are the file specification for both.

i. SNVs (Single Nucleotide variance): The required file format is tab delimited text file with four columns: chromosome, genomic location, reference allele and alternate allele.

```
chr1    100     A       T
chr1    1000    C       T
chr2    100     G       C
……..
```

ii. INDELs: The required file format is a tab limited text file with four columns: chromosome, genomic start, genomic stop and information about the indel. **For insertions**: insertion then “genomic start” = “genomic stop”; **for deletions**: start ≠ stop, stop=start + #Bases. Last column starts with `+` or `-` for insertion or deletion respectively, followed by base/s inserted or deleted followed by supported reads and read depth.

```
chr1    100     100     +A:34/45
chr1    1000    1001    -T:23/34
……..
```

Other sequencing platforms: for all non-Illumina platforms (SOLiD, Roche454, etc.), TREAT only accepts the **called variants** as the input file for the variant annotation. The format is described earlier in the document. Sequence alignment and variant calling are not supported.

Installation

User can download the latest version of TREAT from
http://ndc.mayo.edu/mayo/research/biostat/stand-alone-packages.cfm

- The file to download: TREAT_1.0.1.tar.gz
- Move the file to an appropriate directory (<your_directory>) and run the following command under <your_directory> to un-compress:

  ```
  tar -zxvf TREAT_1.0.1.tar.gz
  ```

  Note that after uncompressing the tar.gz file, a new folder will be created under <your_directory> and named as: TREAT_1.0.1.

- TREAT set up: under the installation directory, run the following command:

  ```
  source ./setup.sh <TREAT_HOME> <EMAIL>
  ```

The two input parameters required for the setup.sh are:
(1) `<TREAT_HOME>` is the complete path to the TREAT_1.0.1 folder: <your_directory>/TREAT_1.0.1.
(2) Users Email address.

The setup script does the following:

- It allows user to create all the scripts to be executable and set all the environment variables required for the TREAT.
- It also creates three configuration files required for the TREAT to run for the example data set and if user wants to run their own data set and then they can go in to modify it.

After Installation, the following directory structure is created automatically:

```
<TREAT_HOME>
        | _ <bin>
              | _ <all the tools>
        | _ <resource>
              | _ <all the references>
        | _ <scripts>
              | _ <sge>
                    | _ <source code for sge mode>
              | _ <non_sge>
                    | _ <source code for non-sge mode>
        | _ <docs>
              | _ <user manual>
```

```
            | _ <files and folder structure description>
            | _ < IGV Setup document>
            | _ < TREAT workflow image>
            | _ < Column Description for variant Reports>
            | _ <amazon cloud tutorial>
            | _ <example configuration files>
      | _ <test_data>
            | _ <fastq>
                  | _ <sample FASTQs>
            | _ <bam>
                  | _ <sample BAMs>
            | _ <variant>
                  | _ <sample variants files>
            | _ <sample _output>
                  | _ <output structure from 'all' module>
      | _ <example>
            | _ <configuration files for example data sets for all four modules>
            | _ <how to run TREAT various modules >
```

The Validation of the Installation: (with the example data set)

The workflow could be run in two different modes namely Sun Grid Engine (SGE) mode (which requires a cluster environment) and single workstation mode.

Single workstation Mode: Under the `<TREAT_HOME>` directory, run the following command to check the validity of the installation:

```
./scripts/non_sge/treat.sh ./example/run_info_all_non_sge.txt
```

SGE mode: if the user has the SGE environment, under the `<TREAT_HOME>` directory, run the following command to check the validity of the installation:

```
./scripts/non_sge/treat.sh ./example/run_info_all_non_sge.txt
```

Upon successful completion of the test run, you will receive an email notification stating that the workflow is completed and results are ready. The results from the test run are stored in the following folder structure:

```
<TREAT_HOME>
      | _ <test_data>
            | _ <LastName_FirstName>
                  | _ <exome>
                        | _ <allModule>
                              | _ <Reports>
                                    | _ <SNV.cleaned_annot.xls>
                                    | _ <SNV.cleaned_annot_filtered.xls>
                                    | _ <INDEL.cleaned_annot.xls>
                                    | _ <INDEL.cleaned_annot_filtered.xls>
                              | _ <Reports_per_Sample>
                                    | _ <sample*.SNV.cleaned_annot.xls>
                                    | _ <sample*.SNV.cleaned_annot_filtered.xls>
                                    | _ <sample*.INDEL.cleaned_annot.xls>
                                    | _ <sample*.INDEL.cleaned_annot_filtered.xls>
                              | _ <Main.Document.html>
                              | _ <igv_session.xml>
```

There are other files and folders which are created using TREAT, but it contains intermediate files useful for tertiary analysis. There is a document named as 'overviewFilesAndFolder.pdf; in the <doc> folder under <TREAT_HOME> which describes each folder and file format.

To view the local sequence alignment using IGV (all rows of the first column of each variant report is hyperlinked to the IGV viewer), you will need to go to the IGV home page at http://www.broadinstitute.org/software/igv/home and download and IGV application and load the **igv_session.xml** file.

Alternatively, in the doc folder:

```
<TREAT_HOME>
        | _ <doc>
```

There is a tutorial which includes different steps to setup IGV (takes less than 5 minutes) and to utilize this feature.

Step by Step Instruction to run TREAT

User needs to prepare two configuration files to run TREAT; One for the sample information and the other for the run information. These files can be located anywhere on the file system as long as they are accessible by the treat.sh shell script. The examples of both files can be found in the <example> folder under <TREAT_HOME> directory.

❖ Run TREAT starting with FASTQ files

Create sample information file:

NOTE:

```
Sample name follows '=' sign and then read1 and read2 are tab separated (specify
the name of the FASTQ files)
```

Option 1: One Paired End sample per lane i.e. one fastq per sample per read

```
sampleA=NameOf_FASTQ_file_Read1ForSampleA    NameOf_FASTQ_file_Read2ForSampleA
sampleB=NameOf_FASTQ_file_Read1ForSampleB    NameOf_FASTQ_file_Read2ForSampleB
...
```

Option 2: One Single End sample per lane i.e. one fastq per sample

```
sampleA=NameOf_FASTQ_file_ReadForSampleA
sampleB=NameOf_FASTQ_file_ReadForSampleB
...
```

Option 3: Multiple Lanes per sample for Paired End

```
sampleA=NameOf_FASTQ_file_Read1ForSampleA        NameOf_FASTQ_file_Read2ForSampleA
sampleA=NameOf_FASTQ_file_NextRead1ForSampleA       NameOf_FASTQ_file_NextRead2ForSampleA
....
```

Option 4: Multiple lanes per sample for Single End

```
sampleA=NameOf_FASTQ_file_ReadForSampleA
sampleA=NameOf_FASTQ_file_NextReadForSampleA
....
```

Create Run information file

The run information file contains parameters that are used by the workflow. User should make sure that the each column identifier should remain same as given in example file, followed by '=' sign. Below is an example of the run information file.

```
TOOL=exome
DATE=5/26/2011
ALIGNER=BWA
SNV_CALLER=SNVmix
PAIRED=1
READLENGTH=100
DISEASE=NONE
VARIANT_TYPE=BOTH
PI=baheti_saurabh
OUTPUT_DIR=/TREAT1.0/test/
EMAIL=baheti.saurabh@mayo.edu
SAMPLENAMES=sampleA:sampleB
TOOL_INFO=/TREAT1.0/example/tool_info_sge.txt
SAMPLE_INFO=/TREAT1.0/example/sample_info_all.txt
ANALYSIS=all
OUTPUT_FOLDER=allModule
CENTER=MAYO
PLATFORM=illumina
GENOMEBUILD=hg18
SAMPLEINFORMATION=There are 2 samples for this study
CHRINDEX=1:2:3:4:5:6:7:8:9:10:11:12:13:14:15:16:17:18:19:20:21:22:X:Y:M
LANEINDEX=1:2
NUM_SAMPLES=2
QUEUE=1-day
INPUT_DIR=/TREAT1.0/test_data/fastq/
```

Description of the identifiers in the run information file:

IDENTIFIER	**Format**	**Description**
TOOL	Exome	To create folder structure
DATE	mm/dd/yyyy	Start date of the analysis
ALIGNER	**BWA**/BOWTIE	Aligner
SNV_CALLER	**SNVmix**/GATK	SNV caller
PAIRED	1/0	1 for PE or 0 for SR
READLENGTH	100	Read length of fastq
DISEASE	Cancer	Name of the disease
QUEUE		Optional, specify if using SGE mode
VARIANT_TYPE	BOTH	Type of variant (BOTH/SNV/INDEL)
PI	lastname_firstname	Name of the person running the TREAT

OUTPUT_DIR		output directory location
INPUT_DIR		path to input directory
EMAIL		Email address
SAMPLENAMES	sampleA:sampleB	Name of the sample same as sample info file ':' seperated
TOOL_INFO		path to tool info file
SAMPLE_INFO		path to sample info file
ANALYSIS	all	
OUTPUT_FOLDER		name for the output folder
CENTER	MAYO/TGEN/BMC/XXX	Where the sequencing is done
PLATFORM	Illumina	Platform information
GENOMEBUILD	hg18	Build of genome
SAMPLEINFORMATION		Free Text for HTML report (information about the samples)
CHRINDEX	1:2:3:4:5:6:7:8:9:10:11:12:13:14:15:16:17:18:19:20:21:22:X:Y:M	(list of chromosomes user need to analyze ':' separated)
LANEINDEX	1:2	(lanes are ':' separated)
NUM_SAMPLES	XX	Number of samples

NOTE: bold ones are our recommendations.

Run TREAT:

For Non SGE mode, run the following command under <TREAT_HOME> directory:

```
./scripts/non_sge/treat.sh <PATH TO RUN INFO>/<run info file>
```

For SGE mode, run the following command under <TREAT_HOME> directory:

```
./scripts/sge/treat.sh <PATH TO RUN INFO>/<run info file>
```

The results of the TREAT can be found in:

```
<OUTPUT_DIR>
        | _ <PI>
              | _ <TOOL>
                | _ <OUTPUT_FOLDER>
                        | _ <realigned_data>
                                | _ <sample>
                                        | _ <sample.igv-sorted.bam>
                                        | _ <sample.igv-sorted.bam.bai>
                                | _ <Reports_per_Sample>
                                        | _ <sample.SNV.cleaned_annot.xls>
                                        | _ <sample.SNV.cleaned_annot_filtered.xls>
                                        | _ <sample.INDEL.cleaned_annot.xls>
                                        | _ <sample.INDEL.cleaned_annot_filtered.xls>
                                | _ <Reports>
                                        | _ <SNV.cleaned_annot.xls>
                                        | _ <SNV.cleaned_annot_filtered.xls>
                                        | _ <INDEL.cleaned_annot.xls>
                                        | _ <INDEL.cleaned_annot_filtered.xls>
                                        | _ <variantLocation_SNVs>
                                        | _ <variantLocation_INDELs>
                                | _ <Main_Document.html>
                                | _ <igv_session.html>


<OUTPUT_DIR>
        | _ <PI>
              | _ <TOOL>
                | _ <OUTPUT_FOLDER>
```

The above structure is created from the information supplied by the user in the run info file. The rest of the folder structure is dependent on the analysis module user is specifying. The folder and

files in the structure above are the output from this module. There are other intermediate folders and files created by TREAT that can be useful for tertiary analysis. The user can run the following command to get rid of the intermediate files:

```
./scripts/sge/cleanspace.sh <full path to OUTPUT_FOLDER>
```

❖ Run TREAT using BAM Files

Create sample information file

NOTE:
Sample name followed by '=' sign (specify the name of the BAM file)

```
sampleA=Name_of_the_BAM_file_forSampleA
sampleB=Name_of_the_BAM_file_forSmapleB
...
```

Create Run information file

The run information file contains parameters that are used by the workflow. User should make sure that the each column identifier should remain same as given in example file, followed by '=' sign.

```
TOOL=exome
DATE=5/26/2011
ALIGNER=BWA
SNV_CALLER=SNVmix
PAIRED=1
READLENGTH=100
DISEASE=NONE
VARIANT_TYPE=BOTH
PI=baheti_saurabh
OUTPUT_DIR=/TREAT1.0/test/
EMAIL=baheti.saurabh@mayo.edu
SAMPLENAMES=sampleA:sampleB
TOOL_INFO=/TREAT1.0/example/tool_info_sge.txt
SAMPLE_INFO=/TREAT1.0/example/sample_info_variant.txt
ANALYSIS=variant
OUTPUT_FOLDER=variantModule
CENTER=MAYO
PLATFORM=illumina
GENOMEBUILD=hg18
SAMPLEINFORMATION=There are 2 samples for this study
CHRINDEX=1:2:3:4:5:6:7:8:9:10:11:12:13:14:15:16:17:18:19:20:21:22:X:Y:M
LANEINDEX=1:2
NUM_SAMPLES=2
QUEUE=1-day
INPUT_DIR=/TREAT1.0/test_data/bam/
```

Description of the identifiers in the run information file:

IDENTIFIER	Format	Description
TOOL	Exome	To create folder structure
DATE	mm/dd/yyyy	Start date of the analysis
ALIGNER		Aligner used to generate BAM
SNV_CALLER	**SNVmix**/GATK	SNV caller
PAIRED	1/0	1 for PE or 0 for SR
READLENGTH	100	Read length of fastq
DISEASE	Cancer	Name of the disease
QUEUE		Optional, specify if using SGE mode
VARIANT_TYPE	BOTH	Type of variant (BOTH/SNV/INDEL)
PI	lastname_firstname	Name of the person running the TREAT
OUTPUT_DIR		output directory location
INPUT_DIR		path to input directory
EMAIL		Email address
SAMPLENAMES	sampleA:sampleB	Name of the sample same as sample info file ':' seperated
TOOL_INFO		path to tool info file
SAMPLE_INFO		path to sample info file
ANALYSIS	Variant	
OUTPUT_FOLDER		name for the output folder
CENTER	MAYO/TGEN/BMC/XXX	Where the sequencing is done
PLATFORM	Illumina	Platform information
GENOMEBUILD	hg18	Build of genome
SAMPLEINFORMATION		Free Text for HTML report (information about the samples)
CHRINDEX	1:2:3:4:5:6:7:8:9:10:11:12:13:14:15:16:17:18:19:20:21:22:X:Y:M	(list of chromosomes user need to analyze ':' separated)
LANEINDEX	1:2	(lanes are ':' separated)
NUM_SAMPLES	XX	Number of samples

NOTE: bold ones are our recommendations.

Run TREAT

For Non SGE mode, run the following command under the <TREAT_HOME> directory:

```
./scripts/non_sge/treat.sh <PATH TO RUN INFO>/<run information file>
```

For SGE mode, run the following command under the <TREAT_HOME> directory:

```
./scripts/sge/treat.sh <PATH TO RUN INFO>/<run information file>
```

The results can be found in:

```
<OUTPUT_DIR>
        | _ <PI>
             | _ <TOOL>
               | _ <OUTPUT_FOLDER>
                     | _ <realigned_data>
                           | _ <sample>
                                  | _ <sample.igv-sorted.bam>
                                  | _ <sample.igv-sorted.bam.bai>
                           | _ <Reports_per_Sample>
                                  | _ <sample.SNV.cleaned_annot.xls>
                                  | _ <sample.SNV.cleaned_annot_filtered.xls>
                                  | _ <sample.INDEL.cleaned_annot.xls>
                                  | _ <sample.INDEL.cleaned_annot_filtered.xls>
```

```
                                    | _ <Reports>
                                          | _ <SNV.cleaned_annot.xls>
                                          | _ <SNV.cleaned_annot_filtered.xls>
                                          | _ <INDEL.cleaned_annot.xls>
                                          | _ <INDEL.cleaned_annot_filtered.xls>
                                          | _ <variantLocation_SNVs>
                                          | _ <variantLocation_INDELs>
                                    | _ <Main_Document.html>
                                    | _ <igv_session.html>


<OUTPUT_DIR>
        | _ <PI>
              | _ <TOOL>
                  | _ <OUTPUT_FOLDER>
```

The above structure is created from the information supplied by the user in the run info file. The rest of the folder structure is dependent on the analysis module user is specifying. The folder and files in the structure above are the output from this module. There are other intermediate folders and files created by TREAT that can be useful for tertiary analysis. The user can run the following command to get rid of the intermediate files:

```
./scripts/sge/cleanspace.sh <full path to OUTPUT_FOLDER>
```

❖ Run TREAT from Called Variants

Create sample information file:

NOTE:

```
Variant identifier (SNV or INDEL) followed by ':'; Sample name follows '=' sign
and then specify the name of the file
```

Option 1: User has both SNVs and INDELs

```
SNV:sampleA=nameOfTheVairantFile
INDEL:sampleA=nameOfTheVariantFile
........
```

Option 2: User has only SNVs

```
SNV:sampleA=nameOfFile
........
```

Option 3: User has only INDELs

```
INDEL:sampleA=nameOfFile
........
```

Create Run information file

The run information file contains parameters that are used by the workflow. User should make sure that the each column identifier should remain same as given in example file, followed by '=' sign. Below is the example of the run information file.

```
TOOL=exome
DATE=5/26/2011
ALIGNER=BWA
SNV_CALLER=SNVmix
PAIRED=1
READLENGTH=100
DISEASE=NONE
VARIANT_TYPE=BOTH
PI=baheti_saurabh
OUTPUT_DIR=/TREAT1.0/test/
EMAIL=baheti.saurabh@mayo.edu
SAMPLENAMES=sampleA:sampleB
TOOL_INFO=/TREAT1.0/example/tool_info_sge.txt
SAMPLE_INFO=/TREAT1.0/example/sample_info_annotation.txt
ANALYSIS=annotation
OUTPUT_FOLDER=annotationModule
CENTER=MAYO
PLATFORM=illumina
GENOMEBUILD=hg18
SAMPLEINFORMATION=There are 2 samples for this study
CHRINDEX=1:2:3:4:5:6:7:8:9:10:11:12:13:14:15:16:17:18:19:20:21:22:X:Y:M
LANEINDEX=1:2
NUM_SAMPLES=2
QUEUE=1-day
INPUT_DIR=/TREAT1.0/test_data/variant/
```

Description of the identifiers in the run information file:

IDENTIFIER	**Format**	**Description**
TOOL	Exome	To create folder structure
DATE	mm/dd/yyyy	Start date of the analysis
ALIGNER		Aligner used to generate variants
SNV_CALLER	**NONE**	
PAIRED	NA	1 for PE or 0 for SR
READLENGTH	NA	Read length of fastq
DISEASE	Cancer	Name of the disease
QUEUE		Optional, specify if using SGE mode
VARIANT_TYPE	BOTH/SNV/INDEL	Type of variant (BOTH/SNV/INDEL)
PI	lastname_firstname	Name of the person running the TREAT
OUTPUT_DIR		output directory location
INPUT_DIR		path to input directory
EMAIL		Email address
SAMPLENAMES	sampleA:sampleB	Name of the sample same as sample info file ':' seperated
TOOL_INFO		path to tool info file
SAMPLE_INFO		path to sample info file
ANALYSIS	annotation	
OUTPUT_FOLDER		name for the output folder
CENTER	NA	Where the sequencing is done
PLATFORM	Illumina	Platform information
GENOMEBUILD	hg18	Build of genome
SAMPLEINFORMATION		Free Text for HTML report (information about the samples)

CHRINDEX	1:2:3:4:5:6:7:8:9:10:11:12:13:14:15:16:17:18:19:20:21:22:X:Y:M	(list of chromosomes user need to analyze ':' separated)
LANEINDEX	1:2	(lanes are ':' separated)
NUM_SAMPLES	XX	Number of samples

NOTE: bold one is required.

Run TREAT

For non-SGE mode, run the following command under the <TREAT_HOME> directory:

```
./scripts/non_sge/treat.sh <PATH TO RUN INFO>/run_info.txt
```

For SGE mode, run the following command under the <TREAT_HOME> directory:

```
./scripts/sge/treat.sh <PATH TO RUN INFO>/run_info.txt
```

The results can be found in

```
<OUTPUT_DIR>
        | _ <PI>
            | _ <TOOL>
              | _ <OUTPUT_FOLDER>
                    | _ <Reports_per_Sample>
                            | _ <sample.SNV.cleaned_annot.xls>
                            | _ <sample.SNV.cleaned_annot_filtered.xls>
                            | _ <sample.INDEL.cleaned_annot.xls>
                            | _ <sample.INDEL.cleaned_annot_filtered.xls>
                    | _ <Main_Document.html>
                    | _ <igv_session.xml>


<OUTPUT_DIR>
        | _ <PI>
            | _ <TOOL>
              | _ <OUTPUT_FOLDER>
```

The above structure is created from the information supplied by the user in the run info file. The rest of the folder structure is dependent on the analysis module user is specifying. The folder and files in the structure above are the output from this module. There are other intermediate folders and files created by TREAT that can be useful for tertiary analysis. The user can run the following command to get rid of the intermediate files:

```
./scripts/sge/cleanspace.sh <full path to OUTPUT_FOLDER>
```

Limitations to the workflow

- Sample names should not start with a number and special character "( ){ }[ ] . , $-" is not permitted.
- If a user have multiple BAMs for a sample, then user need to preprocess the all the BAMs to create a merge BAM for the sample, which can be done using Samtools merge module

```
<TREAT_HOME>/bin/samtools-0.1.12a/samtools merge
Usage:   samtools merge [-nr] [-h inh.sam] <out.bam> <in1.bam> <in2.bam> [...]
```

```
Options:       -n       sort by read names
               -r       attach RG tag (inferred from file names)
               -u       uncompressed BAM output
               -R STR   merge file in the specified region STR [all]
               -h FILE  copy the header in FILE to <out.bam> [in1.bam]
```

- User should make sure the Quality values in the FASTQ are in the Sanger format, if not then user need to convert to Sanger quality values.
- All the reference files are pre-processed, so we assume that user is using all the references provided in the package.

Results Navigation

Tools needed for results visualization

a. IGV
b. Adobe (PDF)
c. EXCEL (2007 for larger data sets)
d. Web browser (IE or Firefox, Safari, Chrome)

Appendix

Column header description for the SNV report

Columns	Column Description	Example
IGV Link	IGV link for variant call, Click on the link will take you to the variant position in IGV	chr1:100089177
Chr	Chromosome Index	chr1
Start	Genomic Position	100089177
dbSNP130	SNV id from dbSNP 130 (a rsID is displayed is alleles are same)	rs2307130
Ref	Reference Allele	A
HapMap_CEU_allele_freq	Allele Frequency for Hapmap CEU samples phase II (caucasian)	G/A,0.492/0.508
1kgenome_CEU_allele_freq	Allele Frequency for 1kgenome CEU samples Release 6 (caucasian)	G/A,0.483/0.517
HapMap_YRI_allele_freq	Allele Frequency for Hapmap YRI samples phase II (Yorubaian)	G/A,0.183/0.817
1kgenome_YRI_allele_freq	Allele Frequency for 1kgenome YRI samples Release 6 (Yorubaian)	G/A,0.212/0.788
HapMap_JPT+CHB_allele_freq	Allele Frequency for Hapmap CEU samples phase II (Japanese and Chinese)	A/G,0.478/0.522
1kgenome_JPT+CHB_allele_freq	Allele Frequency for 1kgneome CEU samples Release 6 (Japanese and Chinese)	A/G,0.474/0.526
Alt	Alternate allele	G
GenotypeClass	SNV class [Homozygous Alternate (AltAlt) or Heterozygous Alternate (AltRef)]	GG
Alt-SupportedReads	Number of Reads supporting Alternate Allele	**min mapping and Base quality 20**
Ref-SupportedReads	Number of Reads supporting Reference Allele	**min mapping and Base quality 20**
ReadDepth	Total Reads stack at the variant position	**min mapping and Base**

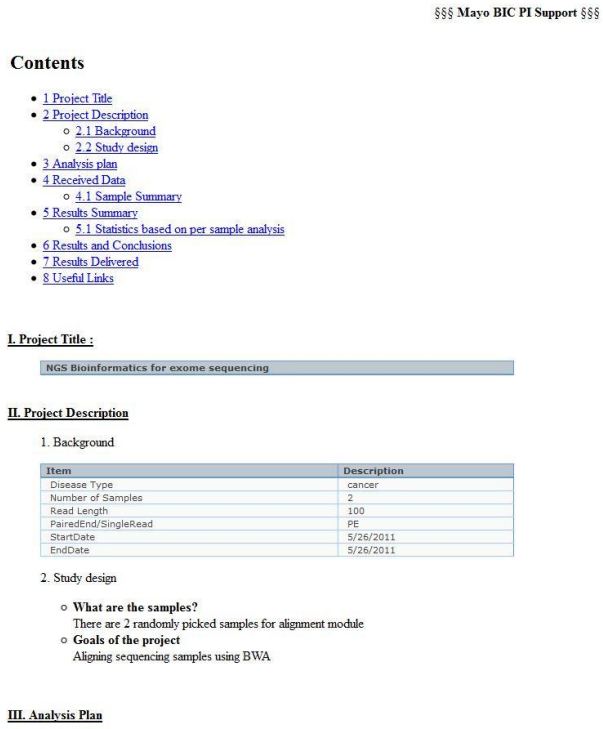
		quality 20
probability/quality	For SNVmix we get posterior probability and For GATK we get quality	**probability > 0.8**
Codons	codon that has been changed, the bases are with respect to + mRNA orientation	-
Transcript ID	Ensemble Transcript ID	ENST00000294724
Protein ID	Ensemble Protein ID	ENSP00000294724
Substitution	Amino Acid substitution with the position information	-
Region	Genomic Region	5' UTR
dbSNP ID	If dbSNP has a variant overlapping at the same position, the rs ID is displayed. However, the alleles may not be the same.	rs2307130:G
SNP Type	Nonsynonymous or Synonymous	-
Prediction	SIFT Prediction (Damaging, Tolerated, DAMAGING *Warning! Low confidence, NA)	Not scored
Score	Ranges from 0 to 1. The amino acid substitution is predicted damaging is the score is <= 0.05, and tolerated if the score is > 0.05	NA
Median Info	Ranges from 0 to 4.32, ideally the number would be between 2.75 and 3.5. This is used to measure the diversity of the sequences used for prediction. A warning will occur if this is greater than 3.25 because this indicates that the prediction was based on closely related sequences	NA
Gene ID	Ensemble Gene ID	ENSG00000162688
Gene Name	Gene Name	AGL
OMIM Disease	OMIM disease from NCBI	GLYCOGEN STORAGE DISEASE III
Average Allele Freqs	CEU Hapmap populations	A,0.60:G,0.40
User Comment	-	-
SynonymousCodonUsage	For every synonymous codon change as shown in the SIFT column 'Codons', this column indicates the percentage occurrence of the codons in Homosapiens. The higher the percentage, the more frequent the codon appears in gene sequences.	-
Difference	This column indicates the difference in percentage between the codon changes. A negative value indicates the synonymous change occurring from a codon with low occurrence rate to a codon with high occurrence rate and vice versa for a positive value.	-
conservation	Score of 1 indicates the variant position overlapping with evolutionary conserved regions in 17/28/44 vertebrates, including mammalian, amphibian, bird, and fish species.	1
Regulation	Score of 1 indicates the variant position overlapping with regulatory potential regions based on short alignment patterns between known regulatory elements and neutral DNA of human, chimpanzee, macaque, mouse, rat, dog, and cow.	1
Tfbs	(TranscriptionFactorBindingSite) Score of 1 indicates the variant position overlapping with transcription factor binding sites conserved in the human/mouse/rat alignment.	0
Tss	(TranscriptionStartSite) Score of 1 indicates the variant position overlapping with transcription start sites (TSS) on the human genome. The TSSs of a gene are important landmarks that help define the promoter regions of a gene.	0

Enhancer	Score of 1 indicates the variant position overlapping distant-acting transcriptional enhancers in the human genome. UCSC couples the identification of evolutionary conserved non-coding sequences with a moderate throughput mouse transgenesis enhancer assay.	0
# inDBSNPOrNot	whether the SNP is in the dbSNP database and/or 1000 Genomes; either echoing the user input in the case of a Maq file, or with no user input, whether represented in these datasets	dbSNP_1000Genomes
Accession	NCBI or CCDS transcript identifier	NM_000644
functionGVS	GVS class of SNP function, using only hg18 and your submitted alleles	splice-3
rsID	dbSNP identifier for SNP	2307130
aminoAcids	list of amino acids for the codon, starting with that of the reference base	none
proteinPosition	the position of the amino acid in the protein, beginning at the N-terminal with the first amino acid at position 1, followed by the total number of amino acids in the protein; the total includes a count for the stop codon	NA
polyPhen	column polyPhen: amino acid substitution impacts	unknown
geneList	Gene Name (USCS)	AGL
Entrez_id	GeneID (UCSC)	178
Gene_title	Gene Descripton (UCSC)	amylo-alpha-1, 6-glucosidase, 4-alpha-glucanotransferase
closest_transcript_id	Splice variant (+/- 2bp)	|NM_000644
Tissue_specificity	Tissue Specifcity information related to the GeneID	Link
Pathway	Pathway information related to the GeneID	link

Column header description for the INDEL report

Columns	Column Description	Examples	
IGV Link	IGV link for variant call,Click on the link will take you to the variant position in IGV	chr1:12862133	chr2:169436315
Chr	Chromosome Index	chr1	chr2
Start	Genomic Start position	12862133	169436315
Stop	Genomic Stoip position	12862133	169436318
Ref	Reference Allele	-	AAA
Alt	Alternate Allele	G	-
Base-Length	Length of INDEL	1	3
Indel-supportedRead	Number of Reads Supporting INDEL	8	0
ReadDepth	Total Reads stack at the INDEL	8	5
# inDBSNPOrNot	whether the SNP is in the dbSNP database and/or 1000 Genomes; either echoing the user input in the case of a Maq file, or with no user input, whether represented in these datasets	none	none
Accession	NCBI or CCDS transcript identifier	NM_001009611	CCDS2229.1
functionGVS	GVS class of SNP function, using only hg18 and your submitted alleles	frameshift	intron
rsID	dbSNP identifier for SNP	0	0
aminoAcids	list of amino acids for the codon, starting with that of the reference base	none	none

proteinPosition	the position of the amino acid in the protein, beginning at the N-terminal with the first amino acid at position 1, followed by the total number of amino acids in the protein; the total includes a count for the stop codon	NA	NA
polyPhen	column polyPhen: amino acid substitution impacts	unknown	unknown
geneList	Gene Name (USCS)	PRAMEF4	SPC25
Entrez_id	GeneID (UCSC)	400735	57405
Gene_title	Gene Descripton (UCSC)	PRAME family member 4	SPC25, NDC80 kinetochore complex component, homolog (S. cerevisiae)
closest_transcript_id	Splice variant (+/- 2bp)		|NM_020675
Tissue_specificity	Tissue Specifcity information related to the GeneID	Link	Link
Pathway	Pathway information related to the GeneID	-	link
		Insertion Example	**Deletion Example**



§§§ Mayo BIC PI Support §§§

Contents

- 1 Project Title
- 2 Project Description
  - 2.1 Background
  - 2.2 Study design
- 3 Analysis plan
- 4 Received Data
  - 4.1 Sample Summary
- 5 Results Summary
  - 5.1 Statistics based on per sample analysis
- 6 Results and Conclusions
- 7 Results Delivered
- 8 Useful Links

I. Project Title :

NGS Bioinformatics for exome sequencing

II. Project Description

1. Background

Item	Description
Disease Type	cancer
Number of Samples	2
Read Length	100
PairedEnd/SingleRead	PE
StartDate	5/26/2011
EndDate	5/26/2011

2. Study design

- **What are the samples?**
  There are 2 randomly picked samples for alignment module
- **Goals of the project**
  Aligning sequencing samples using BWA

III. Analysis Plan

Figure: Sample HTML page

	sampleA	sampleB
Total Reads	2,915,968	2,057,218
Mapped Reads	2,838,835 (97.4 %)	1,999,956 (97.2 %)
Mapped Reads (Q >= 20)	2,608,965 (89.5 %)	1,860,886 (90.5 %)
Used Reads	2,609,035 (89.5 %)	1,860,947 (90.5 %)
Mapped Reads in the Target region	1,564,172 (53.6 %)	1,103,742 (53.7 %)
Called SNVs (SNVmix)	16,468	11,205
SNVs in Target region	9,493	6,745
Transition To Transversion Ratio	2.57	2.63
In dbSNP130	8,750	6,288
NotIn dbSNP130	743	457
Called Indels (GATK)	335	178
Indels in Target region	193	102
Indels leading to frameshift mutations	23	11
Indels in coding regions not frameshift	17	7
Indels in splice sites	2	1
KNOWN VARIANTS(in dbSNP130)		
Total Known SNVs	8,750	6,288
Transition To Transversion Ratio	2.66	2.69
Nonsense	26	22
Missense	2,468	1,888
coding-synonymous	2,482	1,781
coding-notMod3	122	105
Homozygous	3,672	2,645
Heterozygous	5,078	3,643
NOVEL VARIANTS(Notin dbSNP130)		
Total Novel SNVs	743	457
Transition To Transversion Ratio	1.71	2.03
Nonsense	12	4
Missense	261	178
coding-synonymous	127	95
coding-notMod3	12	5
Homozygous	659	408
Heterozygous	84	49

Figure: Sample Statistics table for 'all' module

References

Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25:1754-60. [PMID: 19451168]

Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10:R25.

DePristo, M., Banks, E., Poplin, R., Garimella, K., Maguire, J., Hartl., C., Philippakis, A., del Angel, G., Rivas, M.A, Hanna, M., McKenna, A., Fennell, T. Kernytsky, A., Sivachenko, A, Cibulskis, K., Gabriel, S., Altshuler, D. and Daly, M. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics. 2011 Apr; 43(5):491-498.

Li H.*, Handsaker B.*, Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 25, 2078-9. [PMID: 19505943]

Goya R, Sun MG, Morin RD, Leung G, Ha G, Wiegand KC, Senz J, Crisan A, Marra MA, Hirst M, Huntsman D, Murphy KP, Aparicio S, Shah SP. SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. Bioinformatics. 2010 Mar 15;26(6):730-6.

Contact information

MAYO BIC PI Support
Asif Hossain
Hossain.Asif@mayo.edu
Saurabh Baheti
Baheti.Saurabh@mayo.edu